Efficient network-guided multi-locus association mapping with graph cuts

Chloé-Agathe Azencott*†, Dominik Grimm†, Yoshinobu Kawahara‡ and Karsten M. Borgwardt†§

January 10, 2013

Abstract

As an increasing number of genome-wide association studies reveal the limitations of attempting to explain phenotypic heritability by single genetic loci, there is growing interest for associating complex phenotypes with *sets of genetic loci*. While several methods for multi-locus mapping have been proposed, it is often unclear how to relate the detected loci to the growing knowledge about gene pathways and networks. The few methods that take biological pathways or networks into account are either restricted to investigating a limited number of predetermined sets of loci, or do not scale to genome-wide settings.

We present SConES, a new efficient method to discover sets of genetic loci that are maximally associated with a phenotype, while being connected in an underlying network. Our approach is based on a minimum cut reformulation of the problem of selecting features under sparsity and connectivity constraints, which can be solved exactly and rapidly.

SConES outperforms state-of-the-art competitors in terms of runtime, scales to hundreds of thousands of genetic loci, and exhibits higher power in detecting causal SNPs in simulation studies than existing methods. On flowering time phenotypes and genotypes from *Arabidopsis thaliana*, SConES detects loci that enable accurate phenotype prediction and that are supported by the literature.

1 Introduction

Twin and family/pedigree studies make it possible to estimate the heritability of observed traits, that is to say the amount of their variability that can be at-

*chloe-agathe.azencott@tuebingen.mpg.de

†Machine Learning and Computational Biology Research Group, Max Planck Institute for Developmental Biology & Max Planck Institute for Intelligent Systems Spemannstr. 38, 72076 Tübingen, Germany

‡The Institute of Scientific and Industrial Research (ISIR) Osaka University 8-1 Mihogaoka, Ibaraki-shi, Osaka 567-0047 Japan

§Zentrum für Bioinformatik, Eberhard Karls Universität Tübingen, 72076 Tübingen, Germany

tributed to genetic differences. In the past few years, genome-wide association studies (GWAS), in which several hundreds of thousands to millions of single nucleotide polymorphisms (SNPs) are assayed in up to thousands of individuals, have made it possible to identify hundreds of genetic variants associated with complex phenotypes (Zuk *et al.*, 2012). Unfortunately, while studies associating single SNPs with phenotypic outcomes have become standard, they often fail to explain much of the heritability of complex traits (Manolio *et al.*, 2009). Investigating the joint effects of multiple loci by mapping sets of genetic variants to the phenotype has the potential to help explain part of this missing heritability (Marchini *et al.*, 2005).

While efficient multiple linear regression approaches (Cho *et al.*, 2010; Wang *et al.*, 2011; Rakitsch *et al.*, 2012) make the detection of such multivariate associations possible, they often remain limited in power and hard to interpret. Incorporating biological knowledge into these approaches could help boosting their power and interpretability. However, current methods are limited to predefining a reasonable number of candidate sets to investigate (Cantor *et al.*, 2010; Fridley and Biernacka, 2011; Wu *et al.*, 2011), for instance by relying on gene pathways. They consequently run the risk of missing biologically relevant loci that have not been included in the candidate sets. This risk is made even likelier by the incomplete state of our current biological knowledge. For this reason, our goal here is to use prior knowledge in a more flexible way. We propose to use a biological network, defined between SNPs, to guide a multi-locus mapping approach that is both efficient to compute and biologically meaningful: *We aim to find a small set of SNPs that (a) are maximally associated with a given phenotype and (b) tend to be connected in a given biological network. In addition, this set must be computed efficiently on genome-wide data.* In this paper we assume an additive model to characterize multi-locus association. The network constraint stems from the assumption that SNPs influencing the same phenotype are biologically linked. However, the diversity of the type of relationships that this can encompass, together with the current incompleteness of biological knowledge, makes providing a network in which all the relevant connections are present unlikely. For this reason, while we want to encourage the SNPs to form a subnetwork of the network, we also do not want to enforce that they *must* form a single connected component. Finally, we stress that the method must scale to networks of hundreds of thousands or millions of nodes. Approaches developed to analyze gene networks containing hundreds of nodes, such as that of Nacu *et al.* (2007), Chuang *et al.* (2007) or Li and Li (2008), do therefore not apply.

While our method can be applied to any network between genetic markers, we explore three special types of networks here (see Figure 1):

- *Genomic sequence network* (GS): SNPs adjacent on the genomic sequence are linked together. In this setting we aim at recovering sub-sequences of the genomic sequence that correlate with the phenotype.
- *Gene membership network* (GM): SNPs are connected as in the sequence network described above; in addition, SNPs near the same gene are linked

together as well. Usually, a SNP is considered to belong to a gene if it is either located inside said gene ore within a pre-defined distance of this gene. In this setting we aim more particularly at recovering genes that correlate with the phenotype.

- *Gene interaction network* (GI): SNPs are connected as in the gene membership network described above. In addition, supposing we have a gene-gene interaction network (derived, from example, from protein-protein interaction data or gene expression correlations), SNPs belonging to two genes connected in the gene network are linked together. In this setting, we aim at recovering potential pathways that explain the phenotype.

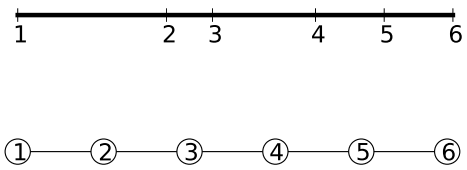
Our task is a feature selection problem in a graph-structured feature space, where the features are the SNPs and the selection criterion should be related to their association with the phenotype considered. Note that our problem is different from subgraph selection problems such as those encountered in chemoinformatics, where each object is a graph and each feature is a subgraph of its own (Tsuda, 2011).

Several approaches have already been developed for selecting graph-structured features. A number of them (Le Saux and Bunke, 2005; Jie *et al.*, 2012) only use the graph over the features to build the learners evaluating their relevance, but do not enforce that the selected features should follow this underlying structure. Indeed they can be applied to settings where the features connectivity varies across examples, while here all individuals share the same network.

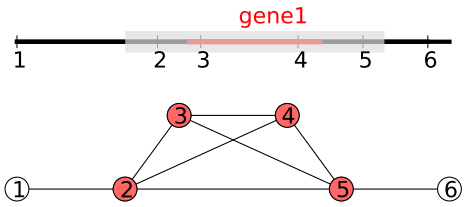
The overlapping group Lasso (Jacob *et al.*, 2009; Liu *et al.*, 2012) is a sparse linear model designed to select features that belong to the union of a small number of predefined groups. If a graph over the features is given, defining those groups as all pairs of features connected by an edge or as all linear subgraphs of a given size yields the so-called graph Lasso. A similar approach is taken by Huang *et al.* (2009): their structured sparsity penalty encourages selecting a small number of base blocks, where blocks are sets of features defined so as to match the structure of the problem. In the case of a graph-induced structure, blocks are defined as small connected components of that graph. As shown in Mairal and Yu. (2011), the overlapping group Lasso mentioned above is a relaxation of this binary problem. As the number of linear subgraphs or connected components of a given size grows exponentially with the number of nodes of the graph, which can reach millions in the case of whole genome SNP data, only the edge-based version of the graph Lasso can be applied to our problem. It is however unclear whether it is sufficient to capture long-range connections between graph nodes.

Li and Li (2008) propose a network-constrained version of the Lasso that imposes the type of graph connectivity we deem desirable. However, their approach has been developed with networks of genes (rather than of SNPs) in mind and does not scale easily to the data sets we envision. Indeed, the implementation they propose relies on a singular value decomposition of the Laplacian of the network, which is intensive to compute and cannot be stored in memory.

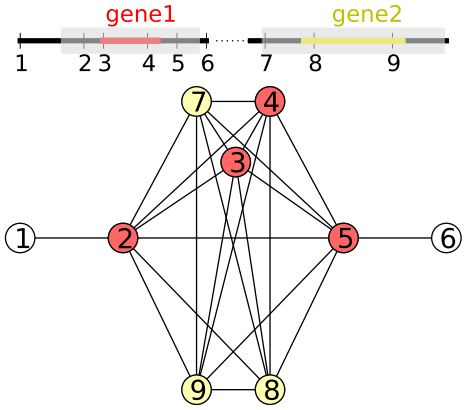
Chuang *et al.* (2007) also searched subnetworks of protein-protein interaction networks that are maximally associated with a phenotype; however, their greedy



(a) Genomic sequence network: SNPs adjacent on the genomic sequence are connected to each other.



(b) Gene membership network: In addition, SNPs near the same gene (i.e., within a specified distance of that gene) are connected.



(c) Gene-interaction network: In addition, SNPs near two interacting genes are connected.

Figure 1: Small examples of the three types of networks considered.

approach requires to fix beforehand a (necessarily small) upper-limit on the size of the subnetworks considered.

In the case of directed acyclic graphs, Mairal and Yu. (2011) propose a minimum flow formulation that make it possible to use for groups (or blocks) the set of all paths of the network. Unfortunately, the generalization to undirected graphs with cycles, such as the SNP networks we consider, requires to randomly assign directions to edges and prune those in cycles without any biological justification. Although this can work reasonably well in practice (Mairal and Yu., 2011), this is akin to artificially removing more than half of the network connections without any biological justification.

In what follows, we formulate the network-guided SNP selection problem as a minimum cut problem on a graph derived from the SNP network in Section 2 and evaluate the performance of our solution both in simulations and on actual *Arabidopsis thaliana* data in Section 3.

2 Methods

2.1 Problem Formulation

Let n be the number of SNPs and m the number of individuals. The SNP-SNP network is described by its adjacency matrix $\boldsymbol{W}$ of size $n \times n$.

A number of statistics based on covariance matrices, such as HSIC (Gretton *et al.*, 2005) or SKAT (Wu *et al.*, 2011), can be used to compte a measure of dependence $\boldsymbol{c} \in \mathbb{R}^n$ between each single SNP and the phenotype. Under the common assumption that the joint effect of several SNPs is additive (which corresponds to using linear kernels in those methods), $\boldsymbol{c}$ is such that the association between a group of SNPs and the phenotype can be quantified as the sum of the scores of the SNPs belonging to this group. In other words, given an indicator vector $\boldsymbol{f} \in \{0,1\}^n$ such that, for any $p \in \{1, \cdots, n\}$, f_p is set to 1 if the p-th SNP is selected and 0 otherwise, the score of the selected SNPs is given by $Q(\boldsymbol{f}) = \sum_{p=1}^{n} c_p f_p = \boldsymbol{c}^\top \boldsymbol{f}$.

We want to find the indicator vector $\boldsymbol{f} \in \{0,1\}^n$ that maximizes the score $Q(\boldsymbol{f})$ while ensuring that the solution is made of connected components of the SNP network. However, in general, it is difficult to find a subset of SNPs that satisfies the above two properties. In fact, given a positive integer k, the problem of finding a connected subgraph with k vertices that maximize the sum of the weights on the vertices, which is equivalent to $Q(\boldsymbol{f})$ of our case, is known to be a strongly **NP**-complete problem (Lee and Dooly, 1996). Therefore, this problem is often addressed based on enumeration-based algorithms, whose runtime grows exponentially with k. To cope with this problem, we consider an approach based on a graph-regularization scheme, which allows us to drastically reduce the runtime.

2.2 Feature Selection with Graph Regularization

Rather than searching through all subgraphs of a given network, we reward the selection of adjacent features through graph regularization. As it is also desirable for biological interpretation and to avoid selecting large number of SNPs in linkage disequilibrium, that the selected sub-networks are small in size, we reward sparse solutions. The first requirement can be addressed by means of a smoothness regularizer on the network (Smola and Kondor, 2003; Ando and Zhang, 2007), while the second one can be enforced with an l_0 constraint:

$$\underset{\boldsymbol{f}\in\{0,1\}^n}{\arg\max} \quad \underbrace{\boldsymbol{c}^\top \boldsymbol{f}}_{\text{association}} \; - \; \underbrace{\lambda \, \boldsymbol{f}^\top \boldsymbol{L} \boldsymbol{f}}_{\text{connectivity}} \; - \underbrace{\eta \; ||\boldsymbol{f}||_0}_{\text{sparsity}} \tag{1}$$

where $\boldsymbol{L}$ is the Laplacian of the SNP network. $\boldsymbol{L}$ is defined as $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$, where $\boldsymbol{D}$ is the diagonal matrix where $\boldsymbol{D}_{p,p}$ is the degree of node p. Note that here, we directly minimize the number of non-zero entries in f and do not require the proxy of an l_1 constraint to achieve sparsity (of course in the case of binary indicators, l_1 and l_0 norms are equivalent).

λ and η are positive parameters that control the importance of the connectedness of selected features and the sparsity regularizer, respectively.

Since $W_{p,q} = 1$ if q is a neighbor of p (also written as $p \sim q$), and 0 otherwise, if we denote by $\mathcal{N}(p)$ the neighborhood of p, then the degree of p can be rewritten $D_{p,p} = \sum_{q\in\mathcal{N}(p)} 1$. The second term in Eq. 1 can therefore be rewritten as

$$\boldsymbol{f}^\top \boldsymbol{L} \boldsymbol{f} = \sum_{p\sim q} (f_p - f_q)^2, \tag{2}$$

and the problem in Eq. (1) is equivalent to

$$\underset{\boldsymbol{f}\in\{0,1\}^n}{\arg\max} \; \sum_{p=1}^{n} f_p(c_p - \eta) - \lambda \sum_{p\sim q} (f_p - f_q)^2 \; . \tag{3}$$

As $(f_p - f_q)^2$ is 1 if $f_p \neq f_q$ and 0 otherwise, it can be seen that the second term penalizes both selecting SNPs that are not connected to one another and selecting only subsets of connected components in the SNP network. Also, as $||\boldsymbol{f}||_0 = \mathbb{1}_n^\top \boldsymbol{f}$ in our case, the third term is equivalent to reducing the individual SNP scores $\boldsymbol{c}$ by a constant $\eta > 0$.

2.3 Min-Cut Solution

Given a graph with vertices $V := \{1, \ldots, n\}$ and adjacency matrix $\boldsymbol{W} \in \mathbb{R}^{n\times n}$, a cut $C(S, V \setminus S)$ ($S \subset \{V \setminus \emptyset\}$) is defined as a partition of the graph. The corresponding cut function $C : \{0,1\}^n \to \mathbb{R}$ is defined as $C(\boldsymbol{f}) = \sum_{p,q=1}^{n} W_{p,q} \boldsymbol{f}_p (1 - \boldsymbol{f}_q)$ where $\boldsymbol{f}_p$ is 1 if $p \in S$ and 0 otherwise. Also, a s/t-cut $C(S, V \setminus S)$ is defined as a cut such that $s \in S$ and $t \in V \setminus S$, where s and t in V are respectively called the *source* and the *sink* of the network. Given a cut $C(S, V \setminus S)$, a set of all pairs (u, v) for $u \in S$ and $v \in V \setminus S$ with positive weight W_{ij} is called the cut-set of

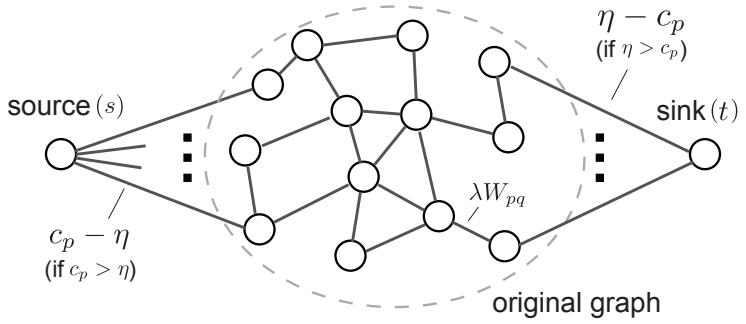


Figure 2: Graph for the s/t-min-cut formulation of the selection of networks of genetic markers.

cut $C(S, V \setminus S)$. Then, for a fixed $s, t \in V$, the problem of finding a s/t-cut that gives a maximum sum of weights on its cut-set is called the s/t min-cut problem. It is well known from the max-flow-min-cut theorem (Papadimitriou and Steiglitz, 1982) that the s/t min-cut problem can be solved efficiently using the maximum-flow algorithm (Goldberg and Tarjan, 1988). In our implementation, we use the Boykov-Kolmogorov algorithm (Boykov and Kolmogorov, 2004). Although its worst case complexity is in $\mathcal{O}(n^2 n_E n_C)$, where n_E is the number of edges of the graph and n_C the size of the minimum cut, it performs much better in practice.

Proposition 1 *Given a graph $\mathcal{G}$ with adjacency matrix $\boldsymbol{W}$, the graph-regularized maximization of score $Q(*)$, i.e., problem* (1)*, is equivalent to a s/t-min-cut for a graph with adjacency matrix $\boldsymbol{M}$ and two additional nodes s and t, where $M_{p,q} = \lambda W_{p,q}$ for $1 \leq p, q \leq n$ and where the weights of the edges adjacent to nodes s and t are defined as*

$$M_{s,p} = \begin{cases} c_p - \eta & \text{if } c_p > \eta \\ 0 & \text{otherwise} \end{cases} \quad \text{and } M_{t,p} = \begin{cases} \eta - c_p & \text{if } c_p < \eta \\ 0 & \text{otherwise} \end{cases} \quad (p = 1, \ldots, n).$$

Proof 1 *The problem in Eq.* (1) *is equivalent to*

$$\underset{\boldsymbol{f} \in \{0,1\}^n}{\arg\min} \; (\eta \mathbb{1}_n - \boldsymbol{c})^\top \boldsymbol{f} + \lambda \boldsymbol{f}^\top \boldsymbol{L} \boldsymbol{f} \,. \tag{4}$$

From Eq. (2)*, we can see that the second term in the above equation is equivalent to a cut function: a pair of nodes (p, q) increases the energy if and only if the nodes are included in different sides of the graph. This is clear because this term can be transformed from the definition of $\boldsymbol{L}$ as*

$$\boldsymbol{f}^\top \boldsymbol{L} \boldsymbol{f} = \sum_{p=1}^{n} f_p \left(D_{p,p} - \sum_{q=1}^{n} W_{p,q} f_q \right) = \sum_{p,q=1}^{n} W_{p,q} f_p (1 - f_q).$$

Moreover, as for the linear term in (1)*, we can see that, if we include the pth-node into* $\mathcal{S}$ *(i.e.,* $f_p = 1$*), the objective increases by* $\eta - c_p$ *if* $\eta > c_p$ *or decreases by* $c_p - \eta$ *if* $c_p > \eta$*. Thus, if we define a vector* $\tilde{\boldsymbol{f}} := [\boldsymbol{f}^\top 1\ 0]^\top$*, we can still represent the objective in Eq.* (1) *as a cut function on graph with adjacency matrix* $\boldsymbol{M}$ *whose entries are defined for* $1 \leq p, q \leq n$ *by* $M_{pq} = \lambda W_{pq}$*,*

$$M_{n+1,p} = \begin{cases} c_p - \eta \text{ if } c_p > \eta \\ 0 \quad \text{otherwise} \end{cases} \text{ and } M_{p,n+2} = \begin{cases} \eta - c_p \text{ if } c_p < \eta \\ 0 \quad \text{otherwise} \end{cases}.$$

Since now the $(n+1)$*-th and* $(n+2)$*-th nodes, which we refer to as* s *and* t*, do not correspond to nodes in the original graph, we can see that this is equivalent to the s/t min-cut problem stated in Proposition 1.* ■

It is therefore possible to use graph cuts to efficiently optimize the objective function defined in Equation 1 and select a small number of connected SNPs maximally associated with a phenotype. We refer to this method as SConES, for Selecting CONnected Explanatory SNPs.

3 Results

We evaluate the ability of SConES to detect networks of trait-associated SNPs on simulated datasets and on datasets from an association mapping study in *Arabidopsis thaliana*.

3.1 Experimental Settings

For all of our experiments, we consider the three SNP networks defined in Section 1: the genomic sequence network, the gene membership network, and the gene interaction network. For SConES, the association term $\boldsymbol{c}$ is derived from Linear SKAT (Wu *et al.*, 2011), which makes it possible to correct for covariates (and therefore population structure).

Linear regression As a baseline for comparisons, we run a linear-regression-based single-SNP search for association, and select those SNPs that are significantly associated with the phenotype (Bonferroni-corrected p-value ≤ 0.05).

Lasso To compare SConES to a method that also considers all additive effects of SNPs simultaneously with a sparsity constraint, but without any network regularization, we also run a Lasso regression (Tibshirani, 1994), using the SLEP implementation (Liu *et al.*, 2009) of the Lasso.

ncLasso In addition, we compare SConES to the network-constrained Lasso ncLasso (Li and Li, 2008), a version of the Lasso with sparsity and graph-

smoothing constraints equivalent to that of SConES. ncLasso solves the following relaxed problem ($\boldsymbol{f} \in \mathbb{R}^n$):

$$\underset{\boldsymbol{f}\in\mathbb{R}^n}{\arg\min}\ \frac{1}{2}\left|\left|\boldsymbol{G}\boldsymbol{f}-\boldsymbol{r}\right|\right|_2^2+\lambda\boldsymbol{f}^\top\boldsymbol{L}\boldsymbol{f}+\eta\left|\left|\boldsymbol{f}\right|\right|_1 \tag{5}$$

We use the reformulation proposed by the authors together with the SLEP implementation Liu *et al.* (2009) of the Lasso. Note that this reformulation requires to compute and store a single value decomposition of $\boldsymbol{L}$ and is therefore not applicable in genome-wide settings where the size of $\boldsymbol{L}$ exceeds $100\,000 \times 100\,000$ by far.

groupLasso and graphLasso Eventually, we also compare our method to the non-overlapping group Lasso (Jacob *et al.*, 2009). The non-overlapping group Lasso solves the following relaxed problem:

$$\underset{\boldsymbol{f}\in\mathbb{R}^n}{\arg\min}\ \frac{1}{2}\left|\left|\boldsymbol{G}\boldsymbol{f}-\boldsymbol{r}\right|\right|_2^2+\lambda\sum_{g\in\mathcal{G}}\left|\left|\boldsymbol{f}^{\mathcal{G}}\right|\right|_2 \tag{6}$$

where $\mathcal{G}$ is a set of (possibly overlapping) predefined groups of SNPs. We consider two versions of the non-overlapping group Lasso:

- graphLasso, for which the groups are directly defined from the same networks as considered for SConES as all pairs of vertices connected by an edge;
- groupLasso, for which the groups are defined sensibly as follows:
  - *Genomic sequence groups* (GS): pairs of adjacent SNPs (note this gives raise to the same groups as for graphLasso with the sequence network);
  - *Gene membership groups* (GM): SNPs near the same gene;
  - *Gene interaction groups* (GI): SNPs near either member of two interacting genes. Here SNPs near genes that are not in the interaction network get grouped by gene.

We use the SLEP implementation (Liu *et al.*, 2009) of the non-overlapping group Lasso, combined with the trick described by Jacob *et al.* (2009) to compute the overlapping group Lasso by replicating features in non-overlapping groups.

Setting the parameters All methods considered, except for the linear regression, have parameters (e.g. λ and η in the case of SConES) that need to be optimized. In our experiments, we run 10-fold cross-validation grid-search experiments over ranges of values of the parameters: 7 values of λ and η each for SConES and ncLasso, and 7 values of the parameter λ for the Lasso and the non-overlapping group Lasso (ranging from 10^{-3} to 10^3). We then pick as

optimal the parameters leading to the most stable selection, defined as that for which the number of SNPs selected in all of the 10 folds is the largest. To avoid trivial solutions picking all or most of the SNPs, we discard results from pairs of parameters that lead to selection of more than 10% of the features.

3.2 Runtime

We first compare the CPU runtime of SConES with that of the linear regression, ncLasso and graphLasso. To assess the performance of our methods, we simulate from 100 to 200 000 SNPs for 200 individuals and generate exponential random networks with a density of 2% between those SNPs.

We report the real CPU runtime of one cross-validation, for set parameters, over a single AMD Opteron CPU (2048KB, 2600MHz) with 512GB of memory, running Ubuntu 12.04 (Figure 3). Across a wide range of numbers of SNPs, SConES is at least two orders of magnitude faster than graphLasso and ncLasso.

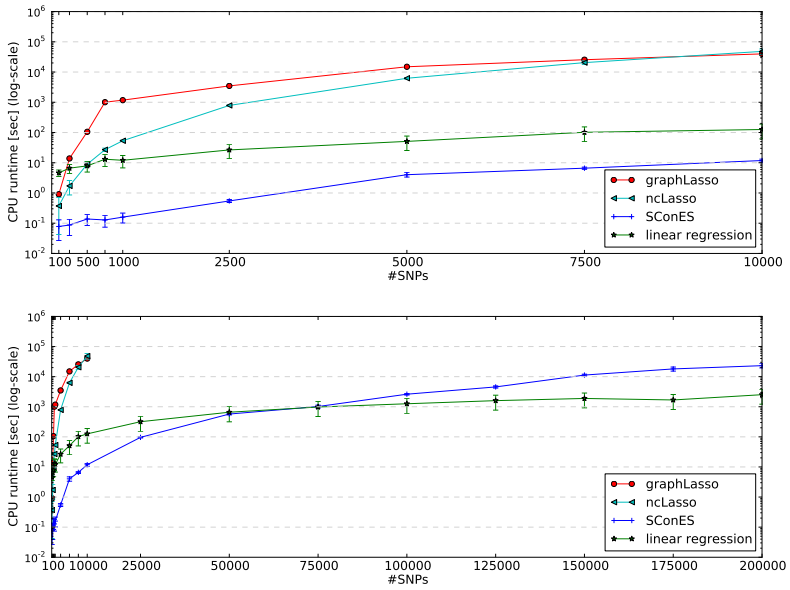


Figure 3: Real CPU runtime comparison between linear regression, ncLasso, non-overlapping group Lasso and SConES, from 100 to 10 000 SNPs (left) and from 100 to 200 000 SNPs (right). After three weeks, ncLasso and non-overlapping group Lasso had not finished running for 25 000 SNPs.

3.3 Simulations

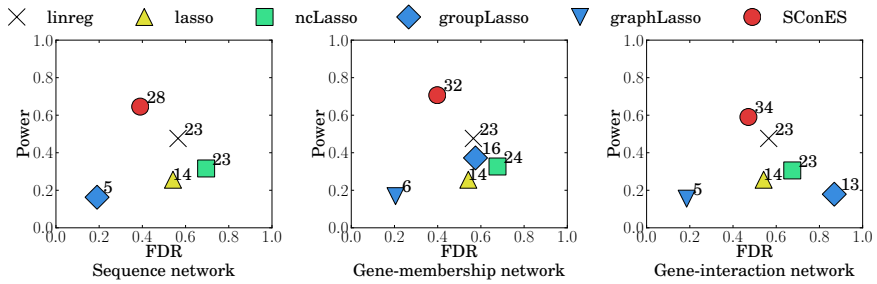
To assess the performance of our methods, we simulate phenotypes for $m = 500$ real *Arabidopsis thaliana* genotypes (214 051 SNPs), chosen at random among those made available by Horton *et al.* (2012), and the *A. thaliana* protein-protein interaction information from TAIR (The Arabidopsis Information Resource, 2012) (resulting in 55 584 646 SNP-SNP connections). We use a window size of 20 000 base-pairs to define proximity of a SNP to a gene, in accordance with the threshold used for the interpretation of GWAS results in Atwell *et al.* (2010). Restricting ourselves to $1,000$ randomly picked SNPs with minor allele frequency larger than 10%, we pick 20 of the SNPs to be causal, and generate phenotypes $y_i = w^\top g_i + \epsilon$, where both the support weights w and the noise ϵ are normally distributed. We consider the following scenarios: (a) the causal SNPs are adjacent in the genomic sequence; (b) the causal SNPs are near the same gene; (c) the causal SNPs are near any of five genes connected in a gene-gene-interaction-network. We then select SNPs using linear regression, Lasso, ncLasso, the two flavors of non-overlapping group Lasso, and SConES as described in Section 3.1. We repeat each experiment 50 times, and compare the selected SNPs of either approach with the true causal ones in terms of power (fraction of causal SNPs selected) and false discovery rate (FDR, fraction of selected SNPs that are not causal).

As SConES returns a binary feature selection rather than a feature ranking, it is not possible to draw FDR curves or compare powers at same FDR as is often done when evaluating such methods. Figure 4 presents the average FDR and power of the different algorithms under the three scenarios, depending on the network used. The closer the FDR-power point representing an algorithm to the upper-left corner, the better this algorithm at maximizing power while mininmizing FDR. As it is easy to get better recall by selecting more SNPs, we also report on the same figure the number of SNPs selected by each algorithm, and show that it remains reasonably close to the true value of 20 causal SNPs. Eventually, we summarize those results with F-scores (harmonic mean of power and one minus FDR) in Table 1.

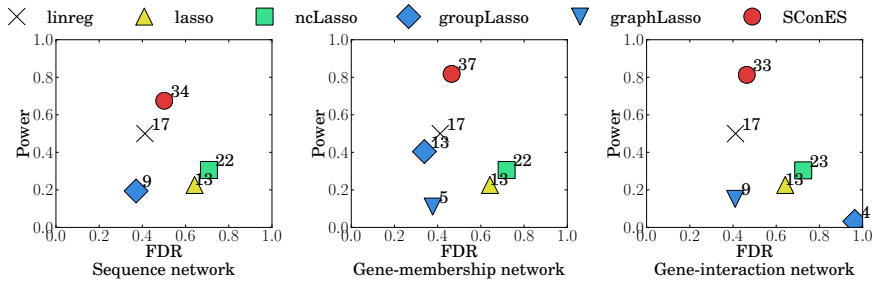
SConES is systematically better than its state-of-the-art comparison partners at leveraging structural information to retrieve the connected SNPs that were causal. While the performance of SConES and ncLasso does depend on the network, the non-overlapping group Lasso is much more sensitive to the definition of its groups. Finally, ncLasso is both slower and less performant than SConES. This indicates that solving the feature selection problem we pose directly, rather than its relaxed version, allows for better recovery of true causal features.

3.4 *Arabidopsis* Flowering Time Phenotypes

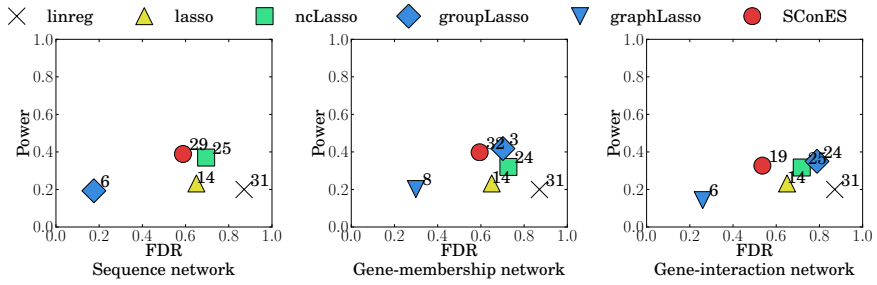
We then apply our method to a large collection of 17 *A. thaliana* flowering times phenotypes from Atwell *et al.* (2010) (up to 194 individuals, 214 051 SNPs). The groups and networks are again derived from the TAIR protein-protein interac-



(a) The true causal SNPs belong to the same genomic segment



(b) The true causal SNPs are near the same gene



(c) The true causal SNPs are near any of five interacting genes

Figure 4: Power and false discovery rate (FDR) of SConES, compared to state-of-the-art Lasso algorithms and a baseline linear regression, in three different data simulation scenarios. Best methods are closest to the upper-left corner. Numbers denote the number of SNPs selected by the method.

Scenario		(a)	(b)	(c)
linear regression		0.45 ± 0.00	0.54 ± 0.00	0.16 ± 0.00
Lasso		0.29 ± 0.02	0.27 ± 0.01	0.28 ± 0.01
ncLasso	GS	0.29 ± 0.02	0.30 ± 0.01	$\mathbf{0.33 \pm 0.01}$
	GM	0.29 ± 0.02	0.29 ± 0.01	0.29 ± 0.01
	GI	0.29 ± 0.02	0.29 ± 0.01	0.29 ± 0.01
groupLasso	GS	0.26 ± 0.01	0.26 ± 0.02	0.29 ± 0.01
	GM	0.39 ± 0.03	0.49 ± 0.02	$\mathbf{0.33 \pm 0.01}$
	GI	0.15 ± 0.03	0.03 ± 0.01	0.26 ± 0.03
graphLasso	GS	0.26 ± 0.01	0.26 ± 0.02	0.29 ± 0.01
	GM	0.25 ± 0.01	0.16 ± 0.01	0.28 ± 0.01
	GI	0.24 ± 0.01	0.20 ± 0.01	0.21 ± 0.01
SConES	GS	$\mathbf{0.55 \pm 0.02}$	0.51 ± 0.02	$\mathbf{0.33 \pm 0.01}$
	GM	$\mathbf{0.56 \pm 0.03}$	$\mathbf{0.59 \pm 0.03}$	0.32 ± 0.02
	GI	0.47 ± 0.02	$\mathbf{0.62 \pm 0.03}$	$\mathbf{0.34 \pm 0.01}$

Table 1: F-scores of SConES, compared to state-of-the-art Lasso algorithms and a baseline linear regression, in three different data simulation scenarios: (a) The true causal SNPs belong to the same genomic segment, (b) The true causal SNPs are near the same gene, and (c) The true causal SNPs are near any of five interacting genes. Best performance in bold.

tion data. We filter out SNPs with a minor allele frequency lower than 10%. We use the first principal components of the genotypic data as covariates to correct for population structure (Price *et al.*, 2006): the number of principal components was chosen by adding them one by one until the genomic control was close to 1.

The direct competitors of SConES on this problem are the methods that also impose graph constraints on the SNPs they select, namely graphLasso and ncLasso. However, they do not scale to datasets such as ours with more than 200k SNPs (see runtime in Figure 3). Hence we had to exclude them from our experiments. Instead, we compare SConES to groupLasso, which uses pairs of neighboring SNPs, SNPs from the same gene or SNPs from interacting genes as pre-defined groups. Note that groupLasso on sequence-neighboring SNPs is identical to graphLasso on the sequence network, which is the only instance of graphLasso whose computation is practically feasible on this dataset. We run Lasso, groupLasso and SConES on the flowering time phenotypes as described in Section 3.1.

To evaluate the quality of the SNPs selected, we perform multivariate linear regression on each phenotype in a cross-validation scheme that uses only the selected SNPs and report its average Pearson's squared correlation coefficient in Figure 5. While the features selected by groupLasso+GS achieve higher predictivity than SConES+GS on most phenotypes, the features selected by SConES+GM are at least as predictive as those selected by groupLasso+GM in two thirds of the phenotypes; the picture is the same for SConES+GI, whose

features are on average more predictive than those of groupLasso+GI.

Next, we checked whether the selected SNPs from the three methods coincide with flowering time genes from the literature. We report in Table 2 the number of SNPs selected by each of the methods and the proportion of these SNPs that are near flowering time candidate genes listed by Segura *et al.* (2012). Here, the picture is reversed: SConES+GS and groupLasso+GI retrieve the highest ratio of SNPs near candidate genes, while groupLasso+GS, SConES+GI and SConES+GM show lower ratios. At first sight, it seems surprising that the methods with highest predictive power retrieve the least SNPs near candidate genes.

To further investigate this phenomenon, we record how many distinct flowering time candidate genes are retrieved on average by the various methods. A gene is considered retrieved if the method selects a SNP near it. Our results are shown in Table 3. Methods retrieving a large fraction of SNPs near candidate genes do not necessarily retrieve the largest number of distinct candidate genes. Good predictive power, as shown in Figure 5, however, seems to correlate with the number of distinct candidate genes selected by an algorithm, not with the percentage of selected SNPs near candidate genes. groupLasso+GI has the highest fraction of candidate gene SNPs among all methods, but detects only three distinct candidate genes. This is probably due to groupLasso selecting entire genes or gene pairs; if groupLasso detects a candidate gene, it will pick most of the SNPs near that gene, which leads to its high candidate SNP ratio in Table 2.

To summarize, SConES is able to select SNPs that are highly predictive of the phenotype. Among all methods, SConES+GM discovers the largest number of distinct genes whose involvement in flowering time is supported by the literature.

4 Discussion and Conclusions

In this article, we defined SConES, a novel approach to multi-locus mapping that selects SNPs that tend to be connected in a given biological network without restricting the search to predefined sets of loci. As the optimization problem of SConES can be solved by maximum flow, our solution is computationally efficient and scales to whole genome data. Our experiments show that our method is about two orders of magnitude faster than the state-of-the-art Lasso-based comparison partners, and can therefore easily scale to hundreds of thousands of SNPs. In simulations, SConES is better at leveraging the structure of the biological network to recover causal SNPs.

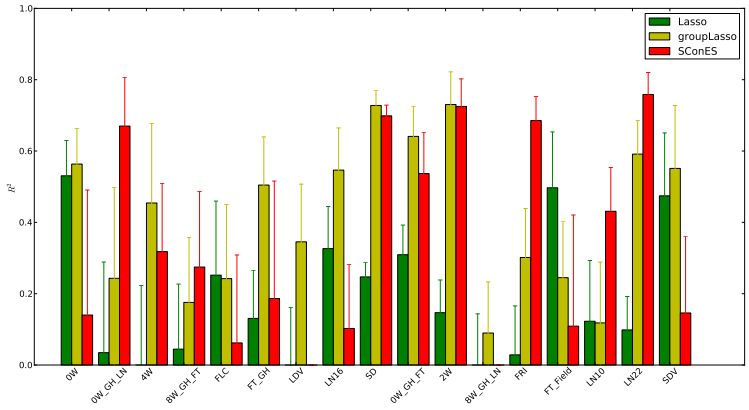
On real GWAS data from *Arabidopsis thaliana*, the predictive ability of the features selected by SConES is superior to that of groupLasso on two of the three network types we consider. When using more biological information (gene membership and gene information), SConES tends to recover more distinct explanatory genes than groupLasso, which in turns leads to better phenotypic prediction.

Phenotype	Lasso	groupLasso			SConES		
		GS	GM	GI	GS	GM	GI
0W	5/85	33/288	59/706	144/547	**123/271**	0/85	0/69
0W GH LN	49/765	13/205	54/478	**128/321**	92/1253	92/1253	92/1253
4W	47/994	23/281	0/166	**80/436**	22/244	0/0	0/0
8W GH FT	10/143	13/166	66/1470	**317/2011**	26/322	26/322	26/322
FLC	**1/31**	2/95	0/101	0/214	0/36	0/2	0/2
FT GH	12/197	37/591	90/841	**177/1417**	0/626	0/59	0/59
LDV	**10/80**	12/100	0/0	0/0	39/674	86/1381	54/1091
LN16	9/222	32/554	138/957	89/1307	**73/73**	0/3	0/4
SD	6/145	36/569	51/863	84/721	**7/59**	**7/59**	**7/59**
0W GH FT	20/194	49/654	52/898	**241/1258**	29/317	29/317	29/317
2W	14/135	42/387	93/610	**126/810**	76/756	78/1185	25/892
8W GH LN	8/122	13/168	0/0	0/0	**11/73**	30/229	20/176
FRI	68/1013	8/64	8/20	**10/10**	101/1274	101/1274	101/1274
FT Field	1/79	7/192	51/221	**52/72**	4/8	4/8	4/8
LN10	21/607	10/184	**18/121**	0/202	165/1921	0/91	0/91
LN22	30/393	68/1132	33/894	81/1023	**140/1378**	**140/1378**	**140/1378**
SDV	4/208	31/588	1/721	105/936	**53/454**	0/8	0/8

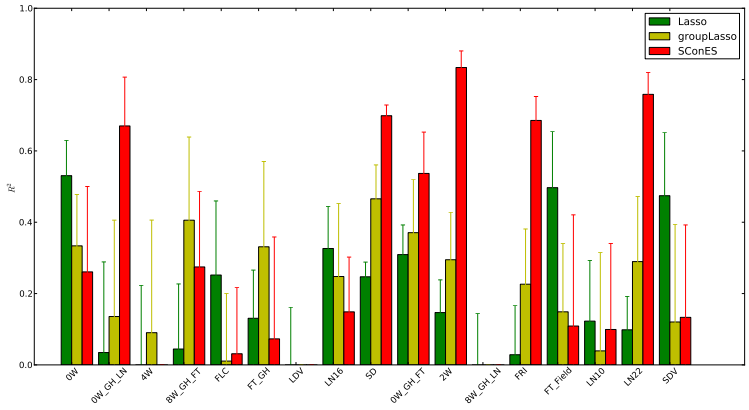
Table 2: Associations detected close to known candidate genes, for all flowering time phenotypes of *Arabidopsis thaliana*. We report the number of selected SNPs near candidate genes, followed by the total number of selected SNPs. Largest ratio in bold. "GS": Genomic sequence network. "GM": Gene membership network. "GI": Gene interaction network.

	Lasso	groupLasso			SConES		
		GS	GM	GI	GS	GM	GI
#SNPs	318	366	533	664	573	450	412
near candidate genes	0.06	0.07	0.09	0.21	0.18	0.08	0.07
candidate genes hit	10	10	1	3	9	12	10

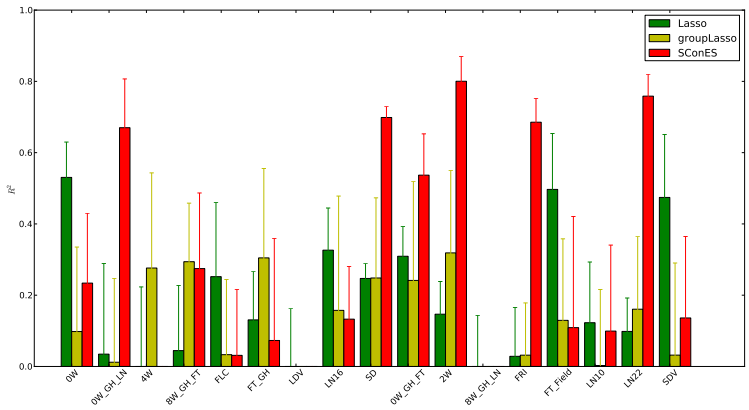
Table 3: Summary statistics, averaged over the *Arabidopsis thaliana* flowering time phenotypes: average total number of selected SNPs ("#SNPs"), average proportion of selected SNPs near candidate genes ("near candidate genes") and average number of different candidate genes recovered ("candidate genes hit") . "GS": Genomic sequence network. "GM": Gene membership network. "GI": Gene interaction network.



(a) Genomic sequence networks



(b) Gene membership networks



(c) Gene interaction networks

Figure 5: Cross-validated predictivity (measured as squared Pearson correlation coefficient between actual phenotype and phenotype predicted by a linear regression over the selected SNPs) of SConES compared to that of the linear regression, Lasso, and groupLasso.

The constraints imposed by groupLasso and SConES are different: while the groups given to groupLasso and the networks passed to SConES come from the same information, the groups force many more SNPs to be selected simultaneously when they may not bring much more information. This gives SConES more flexibility, and makes it less vulnerable to ill-defined groups or networks, which is especially desirable in the light of the current noisiness and incompletedness of biological networks. Our results on the genomic sequence network actually indicate that graphLasso, using pairs of network edges as groups, may achieve the same flexibility as SConES; unfortunately it is too computationally demanding to be run on the most informative networks.

We currently derive the SNP networks from neighborhood along the genome sequence, closeness to a same gene, or proximity to interacting proteins. Refining those networks and exploring other types of networks as well as understanding the effects of their topology and density is one of our next projects.

Let us note that while we do not explicitly consider linkage disequilibrium, the l_0 sparsity constraint of SConES should enforce that when several correlated SNPs are associated with a phenotype, a single one of them is picked. On the other hand, if SConES is given a genomic sequence network such as the one we describe, the graph smoothness constraint will encourage nearby SNPs to be selected together, leading to the selection of sub sequences that are likely to be haplotype blocks. Such a network should therefore only be used when the goal of the experiment is to detect consecutive sequences of associated SNPs.

For now SConES considers an additive model between genetic loci. Future work includes taking pairwise multiplicative effects into account. Replacing the association term in Equation 1 by a sum over pairs of SNPs rather than over individual SNPs results in a maximum flow problem over a fully connected network of SNPs, which cannot be solved straightforwardly, if only because the resulting adjacency matrix is too large to fit in memory on a regular computer. It might be possible, however, to leverage some of the techniques used for two-locus GWAS (Achlioptas *et al.*, 2011; Kam-Thong *et al.*, 2012) to help solve this problem.

Another important extension of SConES is to devise a way to evaluate the statistical significance of the set of selected SNPs. While SConES is speedy enough to allow for permutation tests, the effects of parameter selection on the computation of the p-value and of the optimization on multiple hypotheses testing remain to be evaluated.

Finally, further exciting research topics include applying SConES to larger data sets from human disease consortia, and extending it to the detection of shared networks of markers between multiple phenotypes.

Acknowledgments.

The authors would like to thank Recep Colak, Barbara Rakitsch and Nino Shervashidze for fruitful discussions.

Funding: C.A. is funded by an Alexander von Humboldt fellowship.

References

Achlioptas, P., Schölkopf, B., and Borgwardt, K. (2011). Two-locus association mapping in subquadratic time. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 726–734, New York, NY, USA. ACM.

Ando, R. K. and Zhang, T. (2007). Learning on graph with laplacian regularization. In *Advances in Neural Information Processing Systems 19*.

Atwell, S., Huang, Y. S., Vilhjálmsson, B. J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A. M., Hu, T. T., Jiang, R., Muliyati, N. W., Zhang, X., Amer, M. A., Baxter, I., Brachi, B., Chory, J., Dean, C., Debieu, M., Meaux, J. d., Ecker, J. R., Faure, N., Kniskern, J. M., Jones, J. D. G., Michael, T., Nemri, A., Roux, F., Salt, D. E., Tang, C., Todesco, M., Traw, M. B., Weigel, D., Marjoram, P., Borevitz, J. O., Bergelson, J., and Nordborg, M. (2010). Genome-wide association study of 107 phenotypes in arabidopsis thaliana inbred lines. *Nature*, **465**(7298), 627–631.

Boykov, Y. and Kolmogorov, V. (2004). An experimental comparison of min-cut/max- flow algorithms for energy minimization in vision. *IEEE Trans. Pattern Anal. Mach. Intell.*, **26**(9), 1124 –1137.

Cantor, R. M., Lange, K., and Sinsheimer, J. S. (2010). Prioritizing GWAS results: A review of statistical methods and recommendations for their application. *Am. J. Hum. Genet.*, **86**(1), 6–22.

Cho, S., Kim, K., Kim, Y. J., Lee, J.-K., Cho, Y. S., Lee, J.-Y., Han, B.-G., Kim, H., Ott, J., and Park, T. (2010). Joint identification of multiple genetic variants via elastic-net variable selection in a genome-wide association analysis. *Ann. Hum. Genet.*, **74**(5), 416–428.

Chuang, H.-Y., Lee, E., Liu, Y.-T., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, **3**(140).

Fridley, B. L. and Biernacka, J. M. (2011). Gene set analysis of SNP data: benefits, challenges, and future directions. *Eur J Hum Genet*.

Goldberg, A. V. and Tarjan, R. E. (1988). A new approach to the maximum-flow problem. *Journal of the ACM*, **35**(4), 921–940.

Gretton, A., Bousquet, O., Smola, A., and Schlkopf, B. (2005). Measuring statistical dependence with hilbert-schmidt norms. In *Proceedings of the 16th International Conference on Algorithmic Learning Theory*, pages 63–77. Springer-Verlag.

Horton, M. W., Hancock, A. M., Huang, Y. S., Toomajian, C., Atwell, S., Auton, A., Muliyati, N. W., Platt, A., Sperone, F. G., Vilhjlmsson, B. J., Nordborg, M., Borevitz, J. O., and Bergelson, J. (2012). Genome-wide patterns of genetic variation in worldwide arabidopsis thaliana accessions from the RegMap panel. *Nature Genetics*, **44**(2), 212–216.

Huang, J., Zhang, T., and Metaxas, D. (2009). Learning with structured sparsity. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 417424, New York, NY, USA. ACM.

Jacob, L., Obozinski, G., and Vert, J.-P. (2009). Group lasso with overlap and graph lasso. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 433–440, New York, NY, USA. ACM.

Jie, B., Zhang, D., Wee, C.-Y., and Shen, D. (2012). Structural feature selection for connectivity network-based MCI diagnosis. In P.-T. Yap, T. Liu, D. Shen, C.-F. Westin, and L. Shen, editors, *Multimodal Brain Image Analysis*, volume 7509 of *Lecture Notes in Computer Science*, pages 175–184. Springer Berlin / Heidelberg.

Kam-Thong, T., Azencott, C.-A., Cayton, L., Pütz, B., Altmann, A., Karbalai, N., Sämann, P., Schölkopf, B., Müller-Myhsok, B., and Borgwardt, K. (2012). GLIDE: GPU-Based Linear Regression for Detection of Epistasis. *Hum Hered*, **73**, 220–236.

Le Saux, B. and Bunke, H. (2005). Feature selection for graph-based image classifiers. In J. Marques, N. Perez de la Blanca, and P. Pina, editors, *Pattern Recognition and Image Analysis*, volume 3523 of *Lecture Notes in Computer Science*, pages 147–154. Springer Berlin / Heidelberg.

Lee, H. F. and Dooly, D. R. (1996). Algorithms for the constrained maximum-weight connected graph problem. *Naval Research Logistics*, **43**, 985–1008.

Li, C. and Li, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, **24**(9), 1175–1182.

Liu, J., Ji, S., and Ye, J. (2009). *SLEP: Sparse Learning with Efficient Projections*. Arizona State University.

Liu, J., Huang, J., Ma, S., and Wang, K. (2012). Incorporating group correlations in genome-wide association studies using smoothed group lasso. *Biostat*.

Mairal, J. and Yu., B. (2011). Path coding penalties for directed acyclic graphs. In *Proceedings of the 4th NIPS Workshop on Optimization for Machine Learning (OPT'11)*.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark,

A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F., McCarroll, S. A., and Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, **461**(7265), 747–753.

Marchini, J., Donnelly, P., and Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet*, **37**(4), 413–417.

Nacu, Ş., Critchley-Thorne, R., Lee, P., and Holmes, S. (2007). Gene expression network analysis and applications to immunology. *Bioinformatics*, **23**(7), 850–858.

Papadimitriou, C. H. and Steiglitz, K. (1982). *Combinatorial Optimization: Algorithms and Complexity*. Prentice-Hall Inc., Englewood Cliffs, NJ.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, **38**(8), 904–909.

Rakitsch, B., Lippert, C., Stegle, O., and Borgwardt, K. (2012). A lasso multi-marker mixed model for association mapping with population structure correction. *Bioinformatics*.

Segura, V., Vilhjlmsson, B. J., Platt, A., Korte, A., Seren, ., Long, Q., and Nordborg, M. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat Genet*, **44**(7), 825–830.

Smola, A. and Kondor, R. (2003). Kernels and regularization on graphs. In B. Schölkopf and M. Warmuth, editors, *Learning Theory and Kernel Machines*, volume 2777 of *Lecture Notes in Computer Science*, pages 144–158. Springer Berlin / Heidelberg.

The Arabidopsis Information Resource (2012). TAIR Protein-Protein Interaction. `http://www.arabidopsis.org/portals/proteome/proteinInteract.jsp`.

Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B*, **58**, 267–288.

Tsuda, K. (2011). Graph classification methods in chemoinformatics. In H. H.-S. Lu, B. Schlkopf, and H. Zhao, editors, *Handbook of Statistical Bioinformatics*, Springer Handbooks of Computational Statistics, pages 335–351. Springer Berlin Heidelberg.

Wang, D., Eskridge, K., and Crossa, J. (2011). Identifying qtls and epistasis in structured plant populations using adaptive mixed lasso. *Journal of Agricultural, Biological, and Environmental Statistics*, **16**, 170–184.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, **89**(1), 82–93.

Zuk, O., Hechter, E., Sunyaev, S. R., and Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci. USA*, **109**(4), 1193–1198.